Sample: ERR766214

FASTQ Quality

Filename	ERR766214_R1.fastq.gz	ERR766214_R2.fastq.gz
File Size	366.3 MB	409.2 MB
Q30 Passing	83.25%	62.34%
Mean Read Score	25.5	19.6
Average Read Length	250	250

Read Mapping against ERR766214

Mapped Paired Reads	Mapped Single Reads	Unmapped Reads	Unmapped Percent	Unmapped Assembled Contigs
1,890,407	15,189	17,983	0.9%	assembly not done

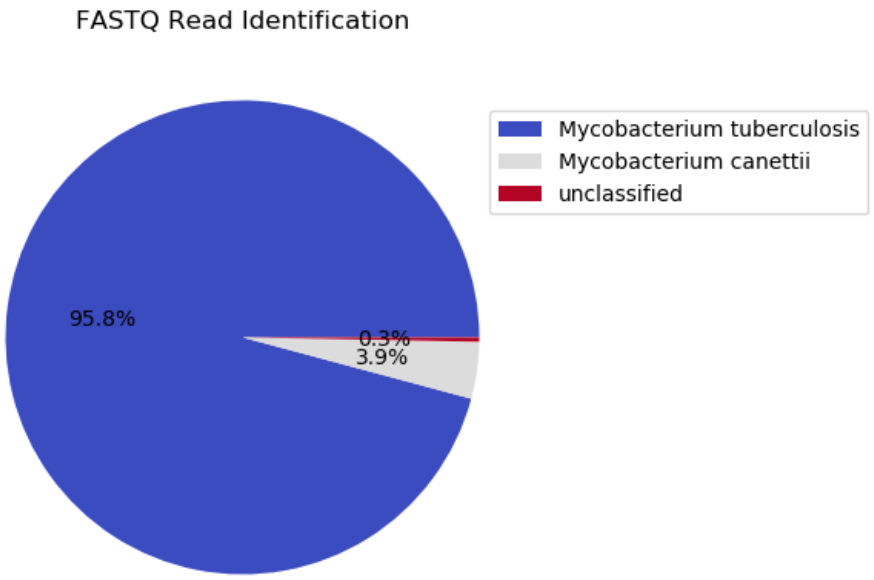
Duplicate Paired Reads	Duplicate Single Reads	Duplicate Percent of Mapped Reads
1,666	5,652	0.2%

BAM File	Reference Length	Genome with Coverage	Average Depth	No Coverage Bases	Quality SNPs
ERR766214_nodup.bam	4,390,827	99.96%	202.8X	1,702	104

Assembly

Contig count	Contig length counts < | 301-999bp | >	Longest contig	Total length	N50	FASTQ calculated mean coverage
249	19 | 176 | 54	478,998	4,390,827	232,241	217.2X

FASTQ Identifications



Identified using: Kraken and Bracken

Genome Size Comparisons

Sourmash Sequence Similarity

Similarity	ID
97.4%	CUWQ01000001.1 Mycobacterium bovis BCG genome assembly bcg_Tice, scaffold BCG_Tice, whole genome shotgun sequence
97.4%	CP002095.1 Mycobacterium bovis BCG str. Mexico, complete genome
97.4%	CP008744.1 Mycobacterium bovis BCG strain 3281, complete genome
97.4%	CUWR01000001.1 Mycobacterium bovis BCG genome assembly bcg_Tokyo, scaffold BCG_Tokyo, whole genome shotgun sequence
97.4%	AM408590.1 Mycobacterium bovis BCG Pasteur 1173P2, complete genome
97.4%	KK355900.1 Mycobacterium tuberculosis BTB04-001 genomic scaffold adOVc-supercont1.1, whole genome shotgun sequence
97.4%	CUWG01000001.1 Mycobacterium bovis BCG genome assembly bcg_China, scaffold BCG_China, whole genome shotgun sequence
97.3%	CUWN01000001.1 Mycobacterium bovis BCG genome assembly bcg_Phipps, scaffold BCG_Phipps, whole genome shotgun sequence
97.3%	CP003900.2 Mycobacterium bovis BCG str. Korea 1168P, complete genome
97.3%	CP014566.1 Mycobacterium bovis BCG str. Tokyo 172 substrain TRCS, complete genome

MLST 2.19.0

Schema-Sequence type: **mycobacteria-268**

MLST Detail: S14Z(17), L35(15), S19(20), L19(19), S12(142), S8(18), L16(22), S7(19)

Results provided by: MLST

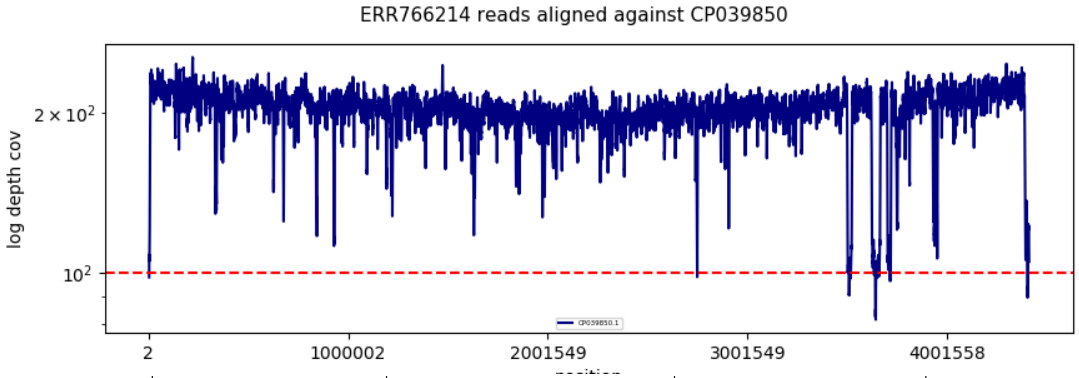
AMRFinderPlus - version 3.11.2, database 2021-03-01.1

AMRFinderPlus Organism Option Not Found: Mycobacterium tuberculosis

Contig ID	Element	Gene	Description	% Cov	% Similar
NODE_1_length_478998_cov_92.920893	AMR	blaC	class A beta-lactamase BlaC	100.0%	100.0%
NODE_23_length_41234_cov_96.839589	AMR	erm(37)	23S rRNA (adenine(2058)-N(6))-methyltransferase Erm(37)	68.72%	100.0%
NODE_3_length_296514_cov_100.171647	AMR	aac(2')-Ic	aminoglycoside N-acetyltransferase AAC(2')-Ic	100.0%	100.0%

Results provided by: AMRFinderPlus

Coverage Graph



ID	Average Depth	Genome Coverage	Genome Length	Ambiguous SNPs
CP039850.1	201.7X	100.0%	4,411,814	51

ID Detail

CP039850.1 Mycobacterium tuberculosis variant bovis strain Danish 1331 chromosome, complete genome

BLAST nt - Assembly Identification	
Nucleotide Representation	Description
2,157,098	CP039850 Mycobacterium tuberculosis variant bovis strain Danish 1331 chromosome, complete genome
578,834	CP110223 Mycobacterium tuberculosis variant bovis strain BCG delta BCG1419c mutant chromosome
515,690	CP039851 Mycobacterium tuberculosis variant bovis strain Danish delta-sapM chromosome, complete geno
294,690	CP003900 Mycobacterium bovis BCG str. Korea 1168P, complete genome
251,894	CP003494 Mycobacterium bovis BCG str. ATCC 35743, complete genome
197,973	CP096841 Mycobacterium tuberculosis variant bovis strain Mb0820 chromosome, complete genome
123,213	CP027035 Mycobacterium tuberculosis variant bovis strain 2002/0476 chromosome, complete genome
63,239	CP085532 Mycobacterium tuberculosis variant bovis BCG strain Moreau PL chromosome, complete genome
59,086	CP002095 Mycobacterium bovis BCG str. Mexico, complete genome
24,133	LR699570 Mycobacterium tuberculosis variant bovis strain Mb3601 genome assembly, chromosome: Mb3601
12,282	LR882500 Mycobacterium tuberculosis variant microti strain Mycobacterium microti 94-2272 genome asse
5,744	CP054014 Mycobacterium tuberculosis strain FDAARGOS_756 chromosome, complete genome
5,513	LC219389 Dioscorea rotundata mitochondrial DNA, contig: TDr_Mt_scaffold16_size5585, cultivar: TDr
4,753	CP054013 Mycobacterium tuberculosis strain FDAARGOS_757 chromosome, complete genome
3,723	CP040832 Mycobacterium tuberculosis variant bovis strain 1 chromosome
2,383	LN714496 TPA_asm: Toxoplasma gondii VEG, chromosome chrVIIa, complete genome
1,838	OX419651 Homo sapiens genome assembly, chromosome: Y
1,588	AC278848 Homo sapiens chromosome 1 clone VMRC53-074E10, complete sequence
1,530	M26007 T.gondii P24 major antigen gene, complete cds, clones Tx(8,10,11)
1,213	AP023783 Homo sapiens DNA, sequence_id: unplaced_0298
1,017	LN714498 TPA_asm: Toxoplasma gondii VEG, chromosome chrVIII, complete genome
890	AP023845 Homo sapiens DNA, sequence_id: unplaced_0360
879	NG_017003 Homo sapiens methyl-CpG binding domain protein 5 (MBD5), RefSeqGene on chromosome 2
874	LN714494 TPA_asm: Toxoplasma gondii VEG, chromosome chrV, complete genome
816	AC078778 Homo sapiens 12 BAC RP11-968A15 (Roswell Park Cancer Institute Human BAC Library) complete

Results provided by: BLAST nt database

Report Description

Input to create this report was either FASTA, or single/paired FASTQ/s.

If built from FASTA, contigs have been identified using Kraken2/Bracken; and assembly metrics, expected genome size compared against total assembly length, MLST, PlasmidFinder, AMRFinderPlus and BLAST results have been reported. In addition to these tests, if report was built from FASTQ/s, sequence technology, FASTQ metrics and coverage graph against the top BLAST hit was also reported. Finally, if Kraken2/Bracken top finding was Salmonella SeqSero2 and CRISPR SeroSeq additionally reported. See table below for summary of tests reported.

Expected results are approximately based on a 4,500,000 base bacterial genome.

When tests are run and there are no significant findings the test's banner and a "no results" message will still be included in this report. If this is not seen, yet test is expected, the report is invalid.

	FASTA	FASTQ	FASTQ Identified as Salmonella
Sequence Technology		X	X
FASTQ metrics		X	X
Kraken2/Bracken	X	X	X
Assembly metrics	X	X	X
Expected Genome Size	X	X	X
MLST	X	X	X
PlasmidFinder	X	X	X
AMRFinderPlus	X	X	X
Coverage Graph		X	X
BLAST	X	X	X
SeqSero2			X
CRISPR SeroSeq			X

Table 1: Test Summary

Sequence Technology

Single FASTQ is assumed to have been generated from Ion S5 System using S5 chemistry. Paired reads are assumed to have been generated from Illumina MiSeq System. Read length is reported based on average read length and rounded up to 50th number.

FASTQ metrics

Filename: Input files
File Size: Compressed input file size (.gz). *Expected: >150MB*
Q30 Passing: Percentage of reads with an average Phred score above 30. *Expected: >70%*
Mean Read Score: Average Phred score of all base calls. *Expected: >30*

Kraken2/Bracken

Bracken computes specie abundance using Kraken2 taxonomy labels. Fraction of reads are reported only when greater than 1% of total reads.

Kraken2 database was built from the following:

Libraries:
archaea: RefSeq complete archaeal genomes
bacteria: RefSeq complete bacterial genomes
plasmid: RefSeq plasmid nucleotide/protein sequences
viral: RefSeq complete viral genomes
human: GRCh38 human genome
fungi: RefSeq complete fungal genomes
plant: RefSeq complete plant genomes
protozoa: RefSeq complete protozoan genomes
UniVec Core: A subset of UniVec chosen to minimize false positive hits to the vector database

Custom genomes:
Bos taurus (cow)
Sus scrofa (pig)
gallus gallus (chicken)
Odocoileus virginianus (white-tailed deer)
Panthera tigris altaica (tiger)
Felis catus (house cat)
Canis lupus familiaris (dog)
Loxodonta africana (elephant)
Tursiops truncatus (bottlenose dolphin)
Anas platyrhynchos (mallard)
Cyprinus carpio (common carp)
Nanorana parkeri (frog)
Aedes albopictus (mosquito)

Unclassified reads were either poor quality or no perfect sample k-mers (k=35) mapped to read sequence from the Kraken2 database build.

Assembly metrics

Reads are assembled using SPAdes when FASTQs are used to generate report. When a FASTA is used SPAdes is assumed assembler.

Metrics can vary greatly. Numbers outside expected values should not always be interpreted as a poor sample,but should be considered against other test results.

Scaffolds: Total contig count of assembly. *Expected: <300*
Total Length: Total length of all sequences. *Expected: 4,500,000*
Longest Scaffold: Longest assembled sequence. *Expected: >200,000*
Scaffolds >1K nt: Sequences greater than 1,000 nucleotides in length. *Expected: >200*
N50: Half of the total assembled length is made from sequence greater than this value. *Expected: >100,000*
Mean Read Depth: Estimated average read depth across total assembly length. *Expected: >50X*

Genome Size Comparisons

Expected genome size is found using top Bracken identification. The genome size is estimated by providing the species name to NCBI which returns the expected size. If NCBI does not return a genome size for the top Bracken identification the Bracken list is iterated from highest to lowest read identification until genome size is found. If an expected genome size is still not found a value of 5,000,000 is used.

The comparison between sample assembly size and expected genome size is very important when determining contamination and genome completeness.

MLST

Assembly file scanned for PubMLST typing schemes using MLST tool.

Schema-Sequence type: Traditional PubMLST typing scheme found
MLST Detail: Gene and allele IDs, see link for explanations for "n", "~", "?", "-" and "n, m" notation.

PlasmidFinder

PlasmidFinder tool used to target plasmids.

Description: NCBI accession number and identification
Length: Sequence length of described accession
Contig ID: Assembly identification of found plasmid
HSP Length: High score pairs length of BLAST database sequence identifying plasmid
% Identity: Percentage of HSP bases similar to matched findings
% Coverage: Percentage of HSP bases aligning to matched findings

AMRFinderPlus

AMRFinderPlus finds acquired antimicrobial resistance genes in assembled nucleotide sequences and resistance-associated point mutations in select taxa. Gene classes found include AMR, point mutations, virulence factors, biocide, heat, acid, and metal resistance genes.

This report includes providing an organism option to AMRFinderPlus if Bracken identification fits select options. When an organism option is provided point mutations are screened and some common findings are suppressed.

AMRFinderPlus Organism Option: Option used to get organism-specific results
Contig ID: Assembly identification of found element
Element: Classification of the AMRFinderPlus gene
Gene: AMR gene
Name: Gene detail
% Cov: Percentage of protein aligning to matched findings
% Similar: Percentage of protein similar to matched findings

Coverage Graph

Reads aligned using BWA to the top BLAST accession hit. When greater than 100X depth-of-coverage log read depth is graphed. If less than 100X a linear read depth is graphed.

ID: NCBI accession aligned against
Average Depth: Average depth-of-coverage across sequence aligned against
Genome Length: Length of genome aligned against

Ambiguous SNPs: Ambiguous SNP count. A high count can indicate poor WGS (i.e. low coverage), a multiple-strain sample or contamination with a closely related strain.
ID Detail: Accession detail

BLAST

Assembly BLAST against nt database. Totaled nucleotide representation identified as top hit.

Nucleotide Representation: Summed contig lengths based on same accession top hit identifications.
Description: NCBI accession number and identification

SeqSero2

SeqSero2 targets the O (wzx and wzy genes) and H (fliC and fljB genes) antigens to determine serotype. When more than one O antigen, or two or more H antigens are found it can indicate the presence of contamination.

Predicted serotype(s): Predicted serotype based on O and H antigens
Predicted antigenic profiles: Shorthand profiles
Predicted subspecies: Predicted subspecies if available
Note: Special notes from SeqSero when provided
Bottom: Contig detail of SeqSero findings